Lecture 20: Concentration inequalities using SDP

Rajat Mittal[*]

IIT Kanpur

Today, we cover another application of optimization in the field of probability. Concentration inequalities bound the amount of deviation from some value of interest by a random variable. The common examples are Markov's inequality and Chebyshev's inequality. The focus will be to come up with such inequalities and their generalization using optimization. Most of the material here is taken from the book of Boyd and Vandenberghe.

1 Probability theory

This section will introduce the notions of probability space and random variables. For a detailed description and definition of these concepts please refer to any standard probability theory textbook.

A *sample space* is the set of outcomes of an experiment. A *probability distribution* is an assignment of probability (a number between 0 and 1) to some collection of subsets (sigma field), such that the total probability is 1. The reader is warned that these definitions are not rigorous and are only intended to remind the intuition behind the actual definitions.

A *random variable* X is a function from the sample space to the set of real numbers. Using the probability distribution on the sample space the sets $(-\infty, x)$ (for $x \in \mathbb{R}$) can be assigned a probability and is known as the distribution function of the random variable.

The expected value of a random variable is the mean of the random variable weighted by probabilities. For a discrete random variable which takes values in range R,

$$\mathbb{E}(X) = \sum_{x \in R} x \ Pr(X = x).$$

If the variable is continuous then integration is used in place of summation.

Concentration inequalities bound the probability of random variable being far away from the quantity of interest (e.g. expected value). The simplest example being Markov's inequality,

$$Pr(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

2 Optimization and concentration inequalities

Suppose $X = \{X_1, \cdots, X_m\}$ is a set of random variables with range $S \subseteq \mathbb{R}^m$. The information is given about these random variables in terms of expectation value of functions $f_i : \mathbb{R}^m \to \mathbb{R}$, say $a_i = \mathbb{E}(f_i)$. We can assume that the expectation of constant function $f_0(x) = 1$ is always given and is equal to 1.

Suppose, we are interested in the probability of the random variable being in some set $C \subseteq S$. The indicator function for set C is the function I_C, s.t., $I_C(x) = 1$ if $x \in C$ and zero otherwise. The probability of X in C is the expected value of this indicator function,

$$Pr(X \in C) = \mathbb{E}(I_C).$$

Taking the same idea as the one to construct a dual solution; if there exist variables y_i such that $\sum_i y_i f_i(x) \geq I_C(x)$ for all $x \in S$, then $\sum_i y_i a_i$ is an upper bound on $\mathbb{E}(I_C)$ and hence on $Pr(X \in C)$.

[*] Thanks to the book of Boyd and Vandenberghe

To get the best upper bound we need to minimize $\sum_i y_i a_i$ where y_i's are "feasible". Hence the following optimization can be made to find the best upper bound,

$$\begin{aligned} \min\ & y_0 a_0 + \textstyle\sum_i y_i a_i \\ \text{s.t. } & y_0 + \textstyle\sum_i y_i f_i(x) \geq 1 \quad \forall x \in C \\ & y_0 + \textstyle\sum_i y_i f_i(x) \geq 0 \quad \forall x \in S \end{aligned} \tag{1}$$

Every feasible solution of the above equation gives a bound on the value of $Pr(X \in C)$. In other words, every feasible solution gives rise to a concentration inequality.

2.1 Example

Consider the case when $m = 1$ and consider two functions f_0, f_1. The first function is the constant function $f_0(x) = 1$ and the second one is identity $f_1(x) = x$. Suppose the expected value of a positive random variable X is μ. We are interested in $Pr(X \geq a)$. The Eqns 1 transform to a linear program,

$$\begin{aligned} \min\ & y_0 + \mu y_1 \\ \text{s.t. } & y_0 + a y_1 \geq 1 \\ & y_0, y_1 \geq 0 \end{aligned} \tag{2}$$

Exercise 1. Show the steps which convert the general Eqns 1 to Eqns. 2 for the special case mentioned above.

Exercise 2. Show that the optimal value of the above linear program is $\min(1, \frac{\mu}{a})$.

So the linear program gives the Markov's inequality with the bonus insight that if $\frac{\mu}{a} \geq 1$ then a better but trivial bound of 1 can be given.

2.2 Another example

Suppose for the set of random variables $X = \{X_1, \cdots, X_m\}$, the first and the second moment are given.

$$\mathbb{E}(X) = a \in \mathbb{R}^m \quad \mathbb{E}(XX^T) = \Sigma \in \mathbb{R}^{m \times m}$$

This means we are given the expected value of every X_i and the expected value of $X_i X_j$ for every pair (i, j). The matrix Σ is symmetric. In this case the linear combination of all the functions can be expressed as ,

$$f(x) = x^T P x + 2q^T x + r.$$

The vector x represents the m random variables and r corresponds to the constant function. Then the expected value of this linear combination can be calculated in terms of Σ and a.

$$\mathbb{E}f(x) = \mathbb{E}(Tr(Pxx^T) + 2q^T x + r) = Tr(\Sigma P) + 2q^T a + r$$

This will be the objective function. For the constraints, $f(x) \geq 0$ and $f(x) \geq 1,\ x \in C$ need to be expressed. We assume that the C is the union of feasible regions of a bunch of inequalities.

$$C = \{x : \exists i,\ \text{s.t. }, a_i^T x \geq b_i\}$$

The constraint $f(x) \geq 0$ is slightly easier. Call $y = \begin{pmatrix} x \\ 1 \end{pmatrix}$.

$$\begin{aligned} f(x) = x^T P x + 2q^T x + r \geq 0 \\ \Leftrightarrow y^T \begin{pmatrix} P & q \\ q^T & r \end{pmatrix} y \geq 0 \\ \Leftrightarrow \begin{pmatrix} P & q \\ q^T & r \end{pmatrix} \succeq 0 \end{aligned}$$

The other condition can be expressed as,

$$a_i^T x \geq b_i \Rightarrow x^T P x + 2q^T x + r \geq 1 \text{ for } i = 1, \cdots, k.$$

This can be equivalently expressed as,

$$\begin{pmatrix} P & q \\ q^T & r-1 \end{pmatrix} \succeq \lambda_i \begin{pmatrix} 0 & \frac{a_i}{2} \\ \frac{a_1^T}{2} & -b_i \end{pmatrix} \text{ for some } \lambda_i \geq 0$$

This can be proved using following theorem. We will not give a proof here (it can be found in the book of Boyd and Vandenberghe, Appendix B.2). The intuition behind this theorem is strong duality and slater's condition.

Theorem 1. *Let M, N be two symmetric matrices, s.t., $x^T M x \geq 0 \Rightarrow x^T N x \geq 0$. If there exist x_0 for which $x_0^T M x_0 > 0$, then*

$$\exists \lambda \geq 0, \text{ such that } N - \lambda M \succeq 0.$$

Hence the upper bound for $Pr(X \in C)$ can be expressed as,

$$\begin{aligned} & \min \; Tr(\Sigma P) + 2q^T a + r \\ \text{s.t.} \quad & \begin{pmatrix} P & q \\ q^T & r \end{pmatrix} \succeq 0 \\ & \begin{pmatrix} P & q \\ q^T & r-1 \end{pmatrix} \succeq \lambda_i \begin{pmatrix} 0 & \frac{a_i}{2} \\ \frac{a_1^T}{2} & -b_i \end{pmatrix} \quad \forall i = 1, \cdots, k \\ & \lambda_i \geq 0 \;\; \forall i = 1, \cdots, k. \end{aligned}$$

Hence the problem of getting an upper bound is converted into a semidefinite program. Notice that it gives a lower bound on the $Pr(X \in P)$, where P is a polyhedra.